VIMSS Computational Microbiology Core
Research on Comparative and Functional Genomics

Adam Arkin[1,2,3] (PI and presenter, aparkin@lbl.gov), Eric Alm[1], Inna Dubchak[1], Mikhail Gelfand[4], Katherine Huang[1], Kevin Keck[1], Frank Olken[1], Vijaya Natarajan[1], Morgan Price[1], Yue Wang[2]

[1]Lawrence Berkeley National Laboratory, [2]University of California, Berkeley, [3]Howard Hughes Medical Institute

The primary roles of the Computational Core are to curate, analyze, and ultimately build models of the data generated by the Functional Genomics and Applied Environmental Microbiology Core groups. The near-term focus of the computational group has been to build the scientific and technical infrastructure necessary to carry out these roles. In particular, the efforts of the computational group have been directed toward three objectives: genomics and comparative genomics, curation and analysis of experimental data from the other core groups, and modeling. Central to each of these goals has been the development of a comprehensive relational database that integrates genomic data and analyses together with data obtained from experiment.

VIMSS DB. At present, well over 100 microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Despite this fact, tools to explore this wealth of information have focused on individual genome sequences. The VIMSS Comparative Genomics database and web-based tools are designed to facilitate cross-species comparison, as well as to integrate experimental data sets with genome-scale functional annotations such as operon and regulon predictions, metabolic maps, and gene annotations according to the Gene Ontology. Over 130 complete genome sequences are represented in the VIMSS Comparative Genomics Database, which is implemented as a MySQL relational database, a Perl library for accessing the database, and a user-friendly website designed for laboratory biologists (http://escalante.lbl.gov). This database is currently being augmented with a novel graph for the efficient query of biological pathways and supporting data. A generic java-based tool for the graphical construction of queries on representations of relational database schema (particular for pathways) in nearly finished and will be applied to VIMSS DB in first quarter 2004..

Web-Based Tools. The VIMSS Comparative Genome Browser allows users to align any number of genomes and identifies predicted orthology relationships between genes. Users can save genes of interest for use in the VIMSS Bioinformatics Workbench (VBW), explore individual genes in depth for information about sequence domains, BLAST alignments, predicted operon structure and functionally related genes inferred from a combination of comparative genomics methods and microarray experiments. The VertiGO comparative gene ontology browser allows users to simultaneously view the genetic complement of any number of genomes according to the Gene Ontology hierarchy. A metabolism browser based on the KEGG metabolic maps allows browsing either the set of enzymes predicted to be present in a single genome, or a comparison highlighting the metabolic differences between two genomes. VBW allows users to create and save lists of genes of interest, and use these lists to investigate phylogenetic relationships by making multiple sequence alignments and phylogenetic trees, as well as apply DNA motif-finding software to identify potential regulatory elements in upstream sequences. Novel motif finding algorithms exploiting the comparative analysis of orthologous proteins have already been accurately difficult motifs such as those from the merR family of regulators of heavy-metal resistance.

Genome Annotation. One of the stated goals of the GTL program is to produce next-generation annotation of target genomes including automated gene functional annotations and prediction of gene regulatory features along with validation of these in silico methods. The most fundamental unit of gene regulation in bacteria is the operon, which is a set of genes that are cotranscribed on a

single RNA transcript. Because few operons have been characterized experimentally outside the model organisms *E. coli* and *B. subtilis*, in silico operon prediction methods have been validated only in these two organisms. We have therefore made accurate and unbiased operon predictions in all bacteria a priority for the computational group. To avoid bias that might arise from using experimental data from only two organisms, we have opted to avoid the use of experimental data entirely using techniques from the field of unsupervised machine learning, and we used gene expression data to estimate the accuracy of our predictions. Key to the success of this approach has been integrating experimental data from the Functional Genomic Core group into our Comparative Genomics Database to validate our in silico procedures. Using our operon prediction tool, we have established that, contrary to reports in the literature, the bacterium *Helicobacter pylori* has a large number of operons. In addition, by examining unusually large non-coding regions within highly conserved operons, we have identified putative pseudogenes in *Bacillus anthracis* that allow us to make phenotypic predictions about the motility of the sequenced Ames strain. As a critical test of our automated genome annotations, we are hosting a genome annotation jamboree in April at the Joint Genome Institute, in which our automated predictions will be verified by human curators. We expect that our annotations, along with confidence levels, will reduce the manual curation workload allowing participants to focus most of their efforts on scientific hypothesis testing.

Functional Genomics. The Functional Genomics Core group is beginning to produce large data sets detailing the response of our target organisms to a variety of stress conditions. The Computational Core group is charged with the responsibility to: store and redistribute these data; assist in the statistical analysis and processing of raw data; and to facilitate comparison of experiments performed with different experimental techniques, different conditions, or different target organisms. As a test case, we have focused most of our efforts in this direction toward gene expression microarray experiments. Among the challenges in the representation of microarray data is developing a data schema that includes both raw and processed data, metadata describing the experimental conditions, and a technical description mapping, for example, each array spot to a corresponding region of the genome sequence and to the set of annotated genes (and their orthologs in other species). We are actively following the development of standards for the representation of this type of data (see Data Management below), and in the meantime have implemented our own simple formats aimed at quick integration with our Comparative Genomics Database. To interpret the results of these experiments, it was necessary to develop a standard set of procedures for data normalization and significance testing and apply it uniformly to raw data from each experiment set, as processed data from different labs commonly involve slightly different analytical techniques. By establishing common methodologies, and a common repository for different experimental results, we were able to meet the goal of facilitating comparative studies as well as using the functional genomic data to test hypotheses generated from our comparative genomic analysis. The methods have been applied to the analysis of pH, salt and heat stress data from *Shewanella oneidensis*. Results from this analysis will be described.

Data Management. During the first year of the project, laboratories in the project began putting in place experimental procedures and are now beginning to produce substantial amounts of data. There is a critical need to define what descriptions of data and experimental procedures (protocols) and factors need to be developed and captured, and to put in place procedures for documenting and recording that information. Recognizing this need, we are in the process of reviewing how experimental procedures are being documented and how experimental factors are being recorded by LBNL affiliated laboratories. This information will be used not only to facilitate information and data acquisition procedures, but also to to enhance and upgrade the BioFiles system for data uploading and the underlying database management system. Working with a consortium of researchers from the wider GTL community we have produced a report on the current status of National Data standards and their advantages and deficiencies and produced a plan for developing standardization of metadata and data representation.